Overview of the DARPA Augmented Cognition Technical Integration Experiment

Mark St. John, David A. Kobus,
Pacific Science & Engineering Group
9180 Brown Deer Road
San Diego, CA 92121
stjohn@pacific-science.com,
dakobus@pacific-science.com

Jeffrey G. Morrison, &
Space and Naval Warfare System Center
53560 Hull Street, Bldg. A33, Rm. 1405
San Diego, CA 92152
jmorriso@spawar.navy.mil

Dylan D. Schmorrow
Defense Advance Research Projects Agency
3701 North Fairfax Drive
Arlington, VA 22203-1714
dschmorrow@darpa.mil

Abstract

The DARPA Augmented Cognition program is developing innovative technologies that will transform human-machine interaction by making information systems sensitive to the capabilities and limitations of the human component of the human-machine system. By taking better advantage of individual human capabilities, and being sensitive to human limitations, it is expected that overall system performance can be improved by an order of magnitude. There have been many recent advances in the field of Cognitive Science toward understanding human decision-making, and the Augmented Cognition program is taking advantage of them in working toward this potential. The technologies developed over the last decade in measuring brain activity and various facets of cognition are serving as the basis for managing the way information is presented to the human operators of complex systems. The Augmented Cognition program is building demonstrable, quantifiable augmentations to human cognitive ability in realistic operational environments. Towards, this goal, the first phase of the Augmented Cognition program was to empirically assess the utility and validity of various psychophysiological measures in dynamically identifying changes in human cognitive activity as decision-makers engaged in cognitive tasks. This report is the culmination of Phase I of the program – *Measuring Cognitive State.* It describes the empirical results of a Technical Integration Experiment (TIE) involving the evaluation of 20 psychophysiologically derived measures (cognitive state gauges) that were developed under Phase I. The gauges came from 11 different research groups, and were developed with a variety of theories and scientific backgrounds. The TIE brought these disparate approaches to assessing cognitive state together to be assessed with a common test protocol using a relatively complex cognitive task that was derived from the real world decision-making requirements seen with tactical decision-makers. This task was developed specifically to meet the needs of assessing these very different gauges with necessary empirical controls, yet still maintain the essential character of those tasks from a cognitive perspective as would be found in an operational command and control environment. Eleven of the gauges successfully identified changes in cognitive activity during the task. This report also describes the integration of individual gauge technologies into suites of gauges to simultaneously measure multiple cognitive indices, and the issues created with sensor technology integration in developing next-generation cognitive state gauges. Additionally, the gauge developers rated the ability of their sensors to integrate with other sensors as fairly high, and most developers reported no problems integrating multiple sensors onto participants. This report summarizes the results from the TIE, and examines the prospects for, and issues that must yet be addressed for, the successful transition of these cognitive state gauges to field-able military person-machine systems in Phase II of the Augmented Cognition program, and beyond.

1 Technical Integration Experiment

The DARPA Augmented Cognition program is developing technologies capable of extending, by an order of magnitude, the information management capacity of war fighters. This will entail selecting from the myriad of theories and sensor technologies related to the measurement of human cognition developed over the last decade, and marrying them with the many advances in automation and information management. For example, a future C^4I (Command, Control, Computers, Communications, and Intelligence) system may assign a task to the specific

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2007**	2. REPORT TYPE	3. DATES COVERED **00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE

Overview of the DARPA Augmented Cognition Technical Integration Experiment

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Pacific Science & Engineering Group,9180 Brown Deer Road,San Diego,CA,92121

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**7**	

operator having the most unused cognitive capacity, or it may filter information or select the mode or style of its presentation based on a particular operator's available capacity to receive information visually, verbally, or through some other sensory modality.

The primary objective for the first phase of the Augmented Cognition program, *Measuring Cognitive State,* was to empirically assess the utility and validity of various psychophysiological measures to dynamically identify changes in human cognitive activity during task performance, and explore potential integration and application issues that would need to be addressed during later phases of the program. Here, we summarize the results of a Technical Integration Experiment (TIE) that provided the culmination for Phase I. This TIE brought together 20 psychophysiological measures (cognitive state gauges) from 11 different research organizations. These measures were demonstrated and assessed in a common test environment that had the complexity and demand characteristics comparable to those seen by a tactical command decision maker.

The gauges used a wide range of sensor technologies, and they were based on very different, yet sometimes overlapping, theoretical approaches. The sensor technologies included functional Near Infra-Red imaging (fNIR), continuous and event-related electrical encephalography (EEG/ERP), eye tracking and pupil dilation, mouse pressure, body posture, heart rate, and galvanic skin response (GSR). Each of the gauges that was evaluated in the study, the type of sensor it used, and the research organization that developed the gauge are listed in Table 1.

Table 1. Summary of Technical Integration Experiment Findings

				Task Load Factors			
Gauge	**Sensor Type**	**Research Group**	**Team**	**Number of Tracks per Wave (6,12,18,24)**	**Track Difficulty (Hi/Lo)**	**Secondary Verbal Task (On/Off)**	**Consistency (% of Participants)**
fNIR							
fNIR (left)	Blood Oxygenation	DrexelU	2	●	○	○	75
fNIR (right)	Blood Oxygenation	DrexelU	2	●	○	○	63
EEG-Continuous							
Percent High Vigilance	EEG	ABM	2	●	○	○	63
Probability Low Vigilance	EEG	ABM	2	●	○	○	75
Executive Load	EEG	QinetiQ/UBristol	3	●	◒	○	100
EEG-ERP							
Motor Effort	ERP-IFF	EGI	1	○	○	○	0
Auditory Effort	ERP-Engage Sound	EGI	1	○	◒	○	0
Loss Perception	ERN-Error Sounds	Sarnoff/Columbia	4	○	○	●	50
Occular-Frontal Source	ERP-Comms	UNewMexico	4	●	○	○	100
Synched Anterior-Posterior	ERP-Comms	UNewMexico	4	○	○	●	100
Visual Source	ERP-Comms	UNewMexico	4	○	○	○	100
Arousal							
Arousal Meter	Inter-Heart Beat Interval	Clemson U	1	○	○	○	0
Arousal	GSR	UHawaii	2	○	○	○	0
Arousal	GSR	AnthroTronix	4	○	○	○	17
Physiological							
Head-Monitor Coupling	Head Posture	UPitt/NRL	1	◒	○	○	43
Head Bracing	Body Posture	UPitt/NRL	1	○	○	○	14
Back Bracing	Body Posture	UPitt/NRL	1	○	○	○	14
Perceptual/Motor Load	Mouse clicks	UHawaii	4	●	●	○	100
Cognitive Difficulty	Mouse pressure	UHawaii	4	●	●	○	100
Index of Cognitive Activity	Pupil dilation	SDSU	floating	◒	○	●	57

Note: Black circles denote statistically significant effects ($p < .05$); half circles denote "marginal" statistical effects ($p < .1$); and White circles denote nonsignificant effects. The final column lists the percentage of participants showing a moderate or high correlation ($r > .3$) between gauge values and the Number of Tracks per Wave. See text for details.

The TIE was not the first attempt to combine multiple psychophysiological technologies and measure cognitive activity during a complex task. For example, Fournier, Wilson, and Swain (1999) used a complex personal-computer-based flight simulation to manipulate user workload while measuring cognitive activity using EEG, heart rate, and eye blinks. Smith, Gevins, Brown, Karnik, and Du (2001) used the same task while measuring EEG, and Van Orden, Limbert, Makeig, and Jung 2001 used a mock air warfare target identification and memory task while measuring eye blinks, fixation durations, and mean pupil diameter. However, the TIE, which required coordinating 11 research groups during simultaneous data collection for 20 gauges, was a major undertaking, and the first attempt to bring so many sensor technologies together at the same time.

For the TIE, the 20 cognitive state gauges were assigned to one of four data collection teams to create suites of gauges that could simultaneously monitor participants as they performed the task. This arrangement was done to: 1) assess compatibility issues among the different gauge technologies, 2) allow the direct comparison of results using the different gauges within a team as they assessed the cognitive state changes of the same participants at the same time, yet 3) allow the use of similar sensor technologies, across teams, that would otherwise compete for access to the same physical locations on test participants.

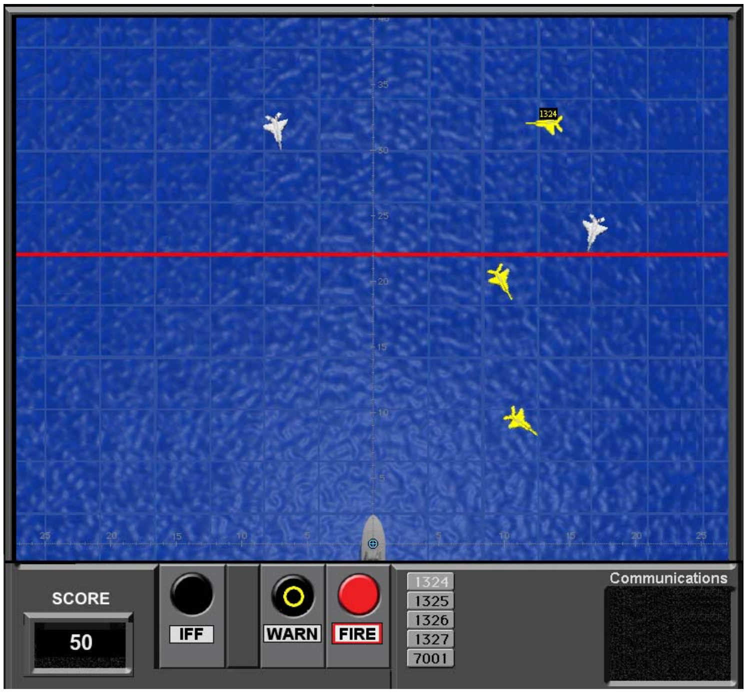

Figure 1. Screen shot of the Airspace Monitoring task in Warship Commander Task

The TIE successfully demonstrated the ability to combine multiple sensors and collect real-time data in a ecologically valid command and control-type decision-making task – which are key requirements of the Augmented Cognition program for transition into Phase II. A key attribute of the TIE was the use of a common experimental test task, under as comparable test conditions as possible across participants and teams. The Warship Commander Task (WCT, St. John, Kobus & Morrison, 2002, see Figure 1) was designed as a basic analog to a Navy air warfare task. This task was based on previous mock air warfare tasks (Ballas, Heitmeyer, & Perez, 1992; Van Orden, Limbert, Makeig, & Jung, 2001), though the pace is faster and the task is more complex in the WCT.

The task was developed to be: 1) suitable for use with undergraduate participants, 2) suitable for stimulating as many aspects of cognition as was feasible, and 3) representative of the complex decision-making environments faced by operational warfighters in tactical command centers. Users performed in a series of 15 minute scenarios during which they monitored a varying number of aircraft (tracks) on a display. They evaluated the tracks and determined if and when it was appropriate to warn them, and if necessary, engage them on the basis of explicit rules of engagement. The task was designed to manipulate a variety of aspects of cognitive activity for the different types of gauges to measure simultaneously, including perception, motor activity, memory, attention, and perceived task load in a semi-realistic command and control-type task.

Figure 2 provides a conceptual illustration of the changing workload demands during the WCT task as perceived by the participants. The pie wedges indicate the proportion of users' activity devoted to each of six dimensions of workload, as defined by the NASA Task Load Index (TLX, NASA-Ames, n.d.; Hart & Staveland, 1988). The left pie chart indicates that during low task load periods of the task, activity on all workload dimensions is low, and users primarily observe and scan the task display. The right pie chart indicates that during high task load periods of the task, temporal and mental demands are high, while other dimensions of workload such as physical demands and frustration remain low, and users have very little time to simply observe the display (The pie wedge proportions are based on previous pilot work). The task, however, did not attempt to explicitly manipulate wakefulness/arousal or physical workload, which has implications for the expected diagnosticity of gauges designed to measure those aspects of cognition.

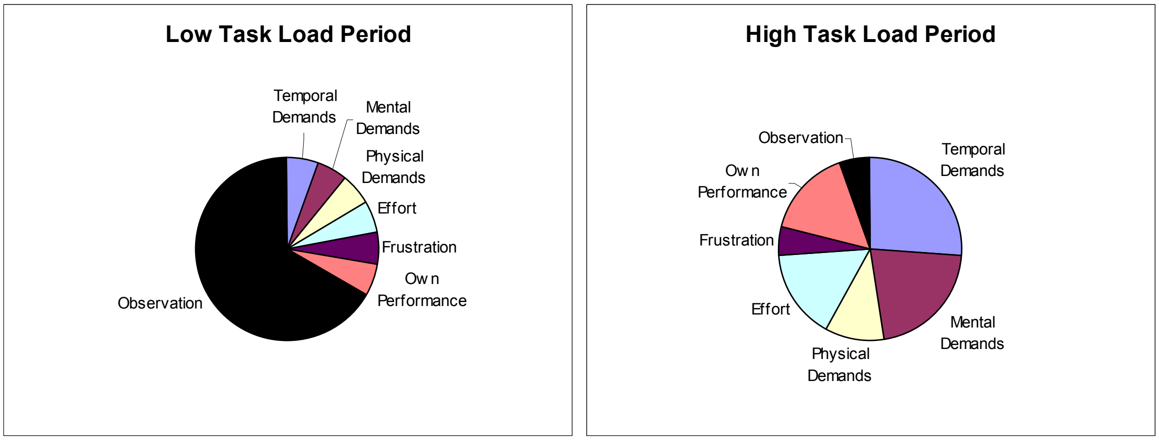

Figure 2. Conceptual illustration of changing workload demands during the WCT task

Cognitive activity, or task load, was varied through three experimental manipulations during the experiment: 1) Number of Tracks per Wave, which varied from 6 to 24 tracks present on the display during each of 12 waves during the course of each scenario, 2) Track Difficulty, which varied between scenarios according to the proportion of potential threat tracks appearing within every wave (High-67% vs. Low-33%) – which required more actions and decisions than other tracks and were thus more complex, and 3) presence or absence of a concurrent secondary auditory/verbal memory task called the Ship Status Task (On or Off) which competed with the primary airspace monitoring task for attentional resources (additional details available at St. John, Kobus, Morrison, & Schmorrow, 2004).

Table 1 summarizes the overall findings of the experiment. For each aspect of task load, a filled black circle in a column indicates that the gauge was statistically sensitive to changes in that specific task load factor (p < .05 according to an analysis of variance). A half-filled black circle indicates that a gauge was "*marginally*" sensitive to changes in that task load factor (p < .10). Given the limitations of sample size and the complexity of the multi-apparatus data collection sessions, as well as the experimental nature of many of the gauges, we felt that reporting these marginal effects was important. At this early stage of development, subtle changes in technology or procedure may dramatically impact the effectiveness of the gauges. An open circle indicates that a gauge was not sensitive to changes in that task load factor.

The final column of Table 1, "Consistency," is an indicator of the consistency with which a gauge detected changes in task load *across* participants. Consistency was measured by first computing the correlation between gauge values and the Number of Tracks per Wave for each scenario for every participant. Then, the mean correlation was computed for each participant. This correlation provided a measure of gauge sensitivity for each participant. Only the Number of Tracks per Wave factor was examined in this analysis since this factor varied from very low task load to very high task load, and many gauges were able to detect changes in it. Finally, the percentage of participants that showed at least a moderately sized mean correlation was computed. A moderately sized mean correlation was defined to be greater than 0.30. These percentages are list in the final column of Table 1.

While some gauges were consistently sensitive for each participant (e.g. Hawaii's mouse-based gauges and QinetiQ's EEG-based gauge), the majority of gauges were sensitive for some participants but not others. It will be important, in future development of these gauges, to determine the sources of variability and attempt to control them.

2 Discussion of Results

Eleven of the gauges successfully correlated with changes in one or more of the task load factors. Two additional gauges showed specific promise for being diagnostic in detecting changes in task load and warrant further development. Since many of the gauges were very early prototypes that were previously unproven, these results are extremely encouraging. In drawing conclusions from these results, it is important to understand several points. First, positive results indicate that a gauge was successful at detecting changes in the factors that were manipulated in the task. It is likely that these gauges will be similarly successful in tasks that have similar attributes and that are measured under comparable environmental conditions. Specifically, tasks that can be characterized as predominantly involving detection, identification, and memory recall (such as computer-based, fast-paced, command and control-type tasks) and that are presented under similar environmental conditions (such as noise, lighting, and time of day), are likely to show comparable results. These gauges may be successful in other types of tasks, as well.

Second, negative results do not necessarily indicate a "failure." The assessment performed during the TIE involved one task, one narrowly defined context, and a relatively small sample size. The data collection environment might have been too noisy for the gauge, or the small sample size might not have contained sufficient statistical power to reveal the sensitivity of a gauge. Furthermore, due to the rapid development of some gauges, the TIE may have been the first attempt to use them on tasks that differed from those used during their development. There also may have been significant individual differences among participants that require the optimization of various sensor technologies and gauge processing algorithms. The assessment of such issues was well beyond the scope of the TIE, or this paper. Consequently, both positive results, and especially, negative results should be interpreted with healthy skepticism.

More importantly, a gauge might be sensitive to aspects of cognition, but not to the specific cognitive task factors that were manipulated by the WCT. For example, in the WCT, the consequences of error are not severe. Further, participants had limited time to acclimate to the myriad combinations of sensors required for the current state of development of some gauges. As a result, it is reasonable to hypothesize that a gauge that measured the stress induced by severe performance anxiety might not react in the WCT, or be sensitive under the necessary test conditions of the TIE. In sum, conclusions from these results must be viewed within the context of the TIE test conditions and the test task; generalization to other tasks and other situations must be drawn with care.

As a class of gauges, the "arousal" gauges stood out for their inability to detect changes in any of the three task load factors. Since arousal gauges are perhaps the best understood of the gauges used during the TIE, their inability to detect changes in cognitive activity during the WCT is somewhat surprising. These results suggest that there may have been a mismatch between the cognitive states measured by these gauges and the cognitive states elicited by the task, or simply that the gauges themselves were insensitive. As noted above, the WCT does not explicitly manipulate stress, arousal, or physical activity other than in terms of mouse and eye movements. Several of the gauge developers suggested that the introduction of stronger negative consequences for errors committed during the task might have produced more measurable stress changes. For example, game score deductions and loud audio error alerts might have created more stress, especially during high task load periods of the task. It may also be the case that well-practiced command and control-type tasks simply do not evoke strong stress responses, and arousal gauges may not be appropriate for measuring changes in workload in such tasks. However, under operational conditions, the negative consequences of errors can be profound, and changes in stress levels may be important to detect. Therefore, we do not recommend eliminating this class of cognitive state gauges at this time.

In either case, the ultimate success of arousal-type gauges will depend on their ability to predict changes in participant performance, rather than changes in arousal, *per se*. It is well known that highly trained operators, such as pilots, can be highly aroused or stressed, for example while landing on an aircraft carrier, with little or no change in their level of arousal, or operational performance (e.g. Berkan, 2000; Menza, 2002). It may be that arousal gauges are better suited for monitoring novices during training and noting how changes in arousal effect human learning. These issues are complex, the research literature is large and varied, and there appear to be many factors that influence the impact of stress on operational performance. More research is required in this area to better understand the relationships between task load, stress, and performance outcomes in different types of command and control tasks and different levels of expertise and motivation.

Another class of gauges, the ERP gauges, showed mixed results: some were effective, while others were not. The development and use of ERP gauges is somewhat problematic in that the user's task must be well understood to identify appropriate task events to measure. It is also necessary to have some means of determining when these events occur during the task. The WCT provided this information to each gauge, but gauges may not have this luxury in real tasks. If these problems can be addressed, then this class of gauges has the potential to measure specific cognitive processing occurring during a task.

The continuous EEG, fNIR, and ICA gauges, on the other hand, all showed substantial promise for detecting changes in workload. For the TIE, they measured average cognitive activity throughout each wave, but it appears quite possible that they could also measure changes in cognitive activity at much finer time scales. Although the EEG gauges, as a group, measured global cognitive functions, such as attention and executive load, there is support for the idea that EEG measures could also be tailored for more specific cognitive processes (Pleydell-Pearce, Whitecross, & Dickson, 2003)

3 Implications

In addition to evaluating the effectiveness of each gauge individually, the TIE evaluated the practical issue of the ability to combine the sensor hardware into useable suites. The gauge "teams" were arranged so that each contained a mix of compatible technologies, although specific assignments were somewhat arbitrary. Overall, all developers rated the ease of integration as fairly high, and most developers reported no problems integrating sensors onto participants. For example, the gauges from Clemson University (arousal) and the University of Pittsburgh/National Research Laboratory (head and body posture) were designed to compliment any other gauge during the TIE. The most common difficulty arose from the lack of headspace available for multiple sensors and the time required to attach and verify their placement. The development of integrated headgear for multiple sensors should be able to address these concerns. The introduction of wireless technology for transmitting sensor data to computers is also highly promising.

From the user's perspective, the TIE experience highlighted the need to make the gauge hardware comfortable, mobile, and convenient enough to gain user acceptance and to become practical for military applications. War fighters cannot be constrained by bulky, uncomfortable equipment that is difficult or tedious to use. Usability is going to be a critical factor in the successful development of augmented cognition systems in relatively stationary command and control center environments, and especially in more mobile environments. Applications where the performer is relatively mobile, such as vehicle operators and soldiers, will be orders of magnitude more daunting in their challenges. Many of the gauge/hardware systems are promising in these regards, but this issue will only increase in its importance as the Augmented Cognition program moves forward to more applied settings.

Another practical concern is the need to understand and address potential sources of electro-magnetic frequency (EMF) interference, both between sensors, various bio-amplifiers and with environmental factors. Several sources of physical and electro-magnetic interference were identified and resolved prior to the TIE data collection event. Other interference was noted on an intermittent basis in the test facility, with no clear source or technical resolution. As we look to the application of these technologies to military environments, it is almost certain that additional sources will appear - operational environments are often noisy and filled with electrical-magnetic interference from many sources. Again, though many improvements in filtering or adapting to this interference have been made, this issue will only grow in importance as augmented cognition becomes a reality.

In sum, the TIE results point to the great potential for a number of psychophysiological gauges to sensitively and consistently detect changes in cognitive state (activity) during relatively complex command and control-type tasks and to their practical integration into an effective sensor suite. Phase I of the Augmented Cognition program has achieved its goal of providing a solid foundation for the demonstration of augmented cognition systems. The primary objective of the TIE was to demonstrate the successful integration of multiple psychophysiological gauges to detect changes in cognitive states in real-time. The goal for Phase II will be to take these gauges and incorporate them into systems for demonstrating the real-time manipulation of cognitive states as the basis for augmenting cognition.

References

Ballas, J. A., Heitmeyer, C. L., & Perez, M. A. (1992). *Direct manipulation and intermittent automation in advanced cockpit*s. Technical Report NRL/FR/5534--92-9375. Naval Research Laboratory, Washington, D. C.

Berkan, M. M. (2000). Performance decrement under psychological stress. Human Performance in Extreme Environments, 5, 92-97.

Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training. *International Journal of Psychophysiology, 31,* 129-145.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam, The Netherlands: Elsevier.

Menza, Lt. M.D. (2002, March). The pucker factor. *Approach*. Retrieved June 27, 2003, from http://www.safetycenter.navy.mil/media/approach/issues/mar02/pucker.htm

NASA-Ames (no date). Task Load Index [TLX] Version 1.0, Users' Manual. Available at http://iac.dtic.mil/hsiac/Products.htm#TLX

Pleydell-Pearce, C.W., Whitecross, S.E., & Dickson, B.T. (2003). Multivariate Analysis of EEG: Predicting cognition on the basis of frequency decomposition, inter-electrode correlation, coherence, cross phase, and cross power. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences.*

Smith, M. E., Gevins, A., Brown, H., Karnik, A., &Du, R. (2001). Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction. *Human Factors, 43,* 366-380.

St. John, M., Kobus, D. A., & Morrison, J. G. (2002). A multi-tasking environment for manipulating and measuring neural correlates of cognitive workload. In *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants*. New York, NY: IEEE. pp 7.10 – 7.14.

St. John, M., Kobus, D. A., Morrison, J. G., & Schmorrow, D. (2004). Overview of the DARPA Augmented Cognition technical integration experiment. *International Journal of Human-Computer Interaction, 17*, 131-149.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors, 43,* 111-121.